The Data Warehouse Lifecycle Toolkit, 2nd Edition

Standaardwerk na tien jaar herzien

Jos van Dongen

Toen ik rondzwierf op de Amazon site stuitte ik op de 2nd edition van de Data Warehouse Lifecycle Toolkit. Eerst had ik niets in de gaten (de 2nd edition van de Toolkit staat gewoon in mijn boekenkast) totdat ik me realiseerde dat het om de 'Lifecycle' Toolkit ging! Na bijna tien jaar is er eindelijk een herziene versie van één van de standaard-werken in BI-land.

Een korte rondgang langs collega's leerde me dat het verschijnen van deze bijgewerkte versie aan eenieders aandacht is ontsnapt. Aangezien dit zelfs gold voor de hoofdredacteur van dit blad leek het ons een goed idee om te kijken wat er nu precies anders is aan de tweede editie en of de ideeën van Kimball de tand des tijds een beetje hebben doorstaan. Laten we voor het gemak met het eerste beginnen en de laatste vraag in de conclusie proberen te beantwoorden.

Omvang

Het eerste wat natuurlijk opvalt aan de tweede editie is de buitenkant omdat het omslag in lijn is gebracht met andere recente BI/DWH uitgaven van Wiley (zie de foto). Wie beter kijkt ziet dat het nieuwe boek ook dunner lijkt, en dat blijkt inderdaad te kloppen. Telde de eerste uitgave een kloeke 771 pagina's, bij de opvolger stopt de teller al bij 636. Een afslankkuur van maar liefst 135 bladzijden dus. Daarbij komt ook nog dat volgens Kimball et al. zo'n 40 procent van het materiaal nieuw of aangepast is. Het gaat dus om een behoorlijke make-over die al begint bij het auteursteam. Laura Reeves staat niet meer op de flap en Joy Mundy en Bob Becker zijn het team komen versterken.
Als de naam Joy Mundy enigszins bekend voorkomt kan dat kloppen; zij is eerste auteur van 'The Microsoft Data Warehouse Toolkit' dat in 2006 uitkwam.
De aanpassingen zitten vooral in het laten vervallen van complete hoofstukken, het herstructureren van de opzet waardoor het boek beter de 'lifecycle' zelf volgt en het compacter weergeven van zaken waarover uitgebreidere teksten in andere boeken te vinden zijn. In elk geval worden de meest basale zaken niet meer uitgelegd. Er wordt simpelweg van uitgegaan dat de lezer bekend is met elementaire zaken als ERD modelleren en de bouwstenen van een BI/DWH-oplossing.

Onderdelen als een 'graduate course on the Internet and Security' zijn (gelukkig) ook geschrapt. Hier zijn veel betere bronnen voor en de ontwikkelingen op dit terrein gaan sneller dan welke drukpers dan ook kan volgen. Op het gebied van ETL is er flink aan de tekst gesleuteld, voornamelijk veroorzaakt door het aparte boek (The Data Warehouse ETL Toolkit) dat Kimball hierover samen met Joe Caserta heeft gepubliceerd. Op basis van dit boek worden in de nieuwe Lifecycle Toolkit 34 ETL-subsystemen benoemd, waarna het volgende hoofdstuk ingaat op het daadwerkelijk ontwerpen en bouwen van het ETL-systeem. Het wekt dan ook geen verbazing dat het ETL-deel van het boek bijna verdubbeld is in omvang (van 53 naar 104 pagina's).

> De kracht van het boek ligt juist in het totaaloverzicht dat geboden wordt

Waar Kimball precies geschrapt heeft is wat lastiger te achterhalen, maar de grootste klap komt hierbij uit onverwachte hoek. Zie hiervoor de paragraaf 'dimensioneel modelleren'.

Terminologie

Op het gebied van technologie heeft de tijd zeker niet stilgestaan, en op dat gebied is de Toolkit weer enigzins aangepast aan de moderne tijd. Alle hot topics als SOA, XML, Master Data Management, Appliances en Massive Parallel Processing worden meegenomen of minimaal genoemd. Dit laatste geldt vooral voor SOA, waarvan alleen gemeld wordt dat het in de lijn der verwachting ligt dat dit steeds vaker voor gaat komen. Hier wordt verder niet dieper op ingegaan. Ook real-time processing, in de eerste uitgave nog onbekend, komt er met zeven pagina's bekaaid af en over een onderwerp als (complex) event processing zult u niets terugvinden in de toolkit. Tot slot heeft Kimball naast de introductie van nieuwe termen een aantal bestaande opnieuw gedoopt. Er wordt dus niet meer gesproken van 'Data Staging' maar van 'ETL' en de kreet 'End User Application' is vervangen door 'BI Application'.

Methodiek

In de basis is de Lifecycle toolkit een complete methode om een datawarehouse-project van A tot Z uit te voeren en daarin is niets wezenlijks veranderd. Eigenlijk is dat jammer, want als er iets duidelijk is geworden de laatste tien jaar is dat de meeste systeemontwikkeltrajecten baat hebben bij een methode die uitgaat van dynamiek, zowel in de omgeving, de techniek als de gebruikerswensen. Iemand als Larissa Moss, auteur van de 'Business Intelligence Roadmap' en een autoriteit op het gebied van BI-project management is daarom al enige tijd bezig om Agile-technieken toe te passen binnen BI-projecten, en ook in eigen land is bijvoorbeeld Sander Hoogendoorn met Agile en BI in de weer. Bij Kimball dus niets van dat al. De lifecycle is ongewijzigd op een paar kleine, maar niet onbelangrijk details na. Zo is het blokje 'BI Application design' wat naar voren gehaald en mag nu gelijk starten met het 'Technical Architecture Design' en 'Dimensional Modeling'. De onderdelen 'Maintenance' en 'Growth' zijn nu twee duidelijk gescheiden blokken, wat een betere weergave vormt van de manier waarop dit in een BI-afdeling georganiseerd wordt of zou moeten worden. Dat wil overigens niet zeggen dat beide onderdelen voldoende zijn uitgewerkt. Het onderdeel 'Growth' (Expanding the DW/BI system) beslaat een magere twaalf pagina's en gaat dan ook totaal niet in op situaties die in de praktijk nog wel eens voor willen komen, zoals bijvoorbeeld het vervangen van een bronsysteem ('maar de rapporten moeten wel gewoon blijven werken!').

Dimensioneel modelleren

De kern van Kimball's verhaal blijft natuurlijk het dimensioneel modelleren. Het zal niemand verbazen dat er geen aandacht is voor alternatieve architecturen of modelleerwijzen als Lindstedt's Data Vault en ook Inmon's DW2.0 wordt volledig genegeerd. Wat Kimball wel heeft gedaan is zich in de vijf pagina's tellende paragraaf 'fables and falsehoods about dimensional modeling' verweren tegen alle kritiek die in de loop der tijd op het dimensioneel model is ontstaan. Verder zijn concepten als mini-dimensies en hybride slowly changing dimension technieken opgenomen, maar verwacht geen uitgebreide cursus dimensioneel modelleren. Dit deel van het boek is namelijk geslonken van 175 in de eerste naar 91 pagina's in de tweede editie, dus wie echt de diepte in wil met alle dimensionele concepten dient de Data Warehouse Toolkit, second edition aan te schaffen. Dat is ook niet erg, de Lifecycle Toolkit is bedoeld als 'handboek soldaat' voor het uitvoeren van projecten, niet als naslagwerk waarin alles wordt uitgediept. Dat is ook meteen het grootste verschil tussen de eerste en de tweede editie. De eerste wilde te veel alles in één zijn, bij de huidige versie ligt de nadruk veel meer op checklists, stappenplannen en een duidelijke werkwijze. De vlag dekt dan ook meer de lading dan bij de eerste editie.

Conclusie

Een eenduidige conclusie over deze editie is wat lastig te formuleren. Enerzijds is er veel werk verzet om het boek aan de eisen van de tijd aan te passen en heeft het team zijn best gedaan om de inhoud van het boek beter op de lifecycle belofte aan te laten sluiten. Anderzijds zijn er de afgelopen tien jaar op de onderdelen projectmethodiek, technologie en modellering meer en ook vooral andere ontwikkelingen geweest die niet in het boek voorkomen. Toch ligt de kracht van het boek juist in het totaaloverzicht dat geboden wordt en het is, samen met de op de Kimball site te vinden templates en checklists, nog steeds (of: opnieuw) een waardevolle informatiebron voor iedereen die met BI of datawarehousing te maken heeft.

The Data Warehouse Lifecycle Toolkit, 2nd Edition
Ralph Kimball, Margy Ross, Warren Thornthwaite, Joy Mundy, Bob Becker
Paperback, 636 pagina's
Wiley
ISBN: 978-0-470-14977-5
Januari 2008

Jos van Dongen
Jos van Dongen (jos@tholis.com) is onafhankelijk adviseur, auteur en spreker.